

L'analisi quantitativa in Istituto Esperienze

Guido Gay

Milano, 16 giugno 2010

Struttura della presentazione

- Esaminare i dati
- La survey su comportamenti e conoscenze in campo alimentare
- Imputazione dei valori mancanti
- Indici sintetici
- Riporto all'universo, quando?

Esaminare i dati

Tabella 3.13 - Prospetto di sintesi dei dati finanziari dei progetti attivati nel periodo 2001-2009, suddiviso per comparto

Comparto	Progetti attivati	Valore progetti		Spesa regionale		Compartecipazione %
		euro	%	euro	%	
ZOOTECNICO E FORAGGICOLTURA	61	18.523.452,79	27,82	8.746.817,02	21,86	52,78
PRODUZIONI DI ORIGINE ANIMALE	27	6.068.794,16	9,11	3.526.998,74	8,81	41,88
ORTICOLO	14	3.017.400,52	4,53	1.938.103,94	4,84	35,77
GRANDI COLTUREERBACEE	33	7.688.672,36	11,55	4.675.268,68	11,68	39,19
VITICOLO ENOLOGICO	29	3.519.905,30	5,29	2.065.628,38	5,16	41,32
FRUTTICOLO	25	2.398.390,34	3,60	1.447.591,52	3,62	39,64
FLOROVIVAISMO E COLTURE OFFICINALI	12	2.203.767,17	3,31	1.502.877,76	3,76	31,80
FORESTA - LEGNO	29	4.066.275,09	6,11	2.454.369,69	6,13	39,64
ITTICO E FAUNISTICO VENATORIO	25	1.830.844,41	2,75	1.380.982,64	3,45	24,57
TERRITORIO E AMBIENTE	74	14.128.469,32	21,22	9.744.507,35	24,35	31,03
ANALISI SOCIO - ECONOMICHE	40	3.148.936,97	4,73	2.528.431,57	21,86	19,71
TOTALE 2001 - 2009	369	66.594.908,43	100	40.011.577,29	100	39,92

Tabella 3.12 - Prospetto di sintesi dei dati finanziari dei progetti attivati nel periodo 2001-2009

Anno	Progetti attivati	Valore progetti €	Spesa regionale €	Compartecipazione %
2001	43	4.908.425,26	4.062.468,84	17,23
2002	52	5.796.621,29	4.553.826,62	21,44
2003	62	7.131.258,25	4.981.551,28	30,14
2004	44	7.488.980,48	4.775.826,63	36,23
2005	44	8.113.402,98	4.574.440,10	43,62
2006	42	6.835.695,16	3.960.382,33	42,06
2007	30	6.246.636,59	3.627.853,07	41,92
2008	28	8.558.135,58	5.176.400,61	39,51
2009	24	11.515.752,84	4.298.827,81	62,67
TOTALE 2001 - 2009	369	66.594.908,43	40.011.577,29	39,92

La survey su conoscenze e comportamenti in campo alimentare

*La survey su comportamenti e
conoscenze in campo alimentare*

Rilevazione telefonica dei **comportamenti e conoscenze in campo nutrizionale**, realizzata dal servizio CATI dell'IRER tra il 18 novembre ed il 23 dicembre 2009.

La popolazione obiettivo dell'indagine era costituita dai cittadini maggiorenni della Lombardia.

Campione stratificato per classi di età (18-29 anni; 30-59; 60-74; 75 e oltre) e sesso.

Tra le variabili rilevate il peso e l'altezza, nonché le risposte a dodici quesiti volti ad accertare le conoscenze nutrizionali .

*La survey su comportamenti e
conoscenze in campo alimentare*

**Tabella 6.1 – Universo: popolazione residente al 1 gennaio 2008
per classe di età e sesso (Regione Lombardia)**

Età	Maschi	Femmine	Totale
18-29	600.618	573.024	1.173.642
30-59	2.231.742	2.163.285	4.395.027
60-74	753.268	852.066	1.605.334
75 e oltre	305.822	570.253	876.075
Totale	3.891.450	4.158.628	8.050.078

Fonte: elaborazione IReR su dati ISTAT

*La survey su comportamenti e
conoscenze in campo alimentare*

Tabella 6.4 – Campione effettivo

Età	Maschi	Femmine	Totale
18-29	60	86	146
30-59	217	329	546
60-74	95	109	204
75 e oltre	38	70	108
Totale	410	594	1.004

n. n.

Tabella 6.3 – Campione teorico

Età	Maschi	Femmine	Totale
18-29	75	71	146
30-59	277	269	546
60-74	94	106	199
75 e oltre	38	70	108
Totale	483	517	1.000

*La survey su comportamenti e
conoscenze in campo alimentare*

D12

Quali fra queste tipologie alimentari è particolarmente ricca di Grassi saturi?

- Carne
- **Burro**
- Patate
- Non sa / Non risponde

12 domande sulle conoscenze nutrizionali. Le risposte ricodificate in “giusta” (valore 1) e “sbagliata” (valore 0)

Imputazione valori mancanti

*La survey su comportamenti e
conoscenze in campo alimentare*

¹ L'Indice di Massa Corporea (IMC) è un dato biometrico, utilizzato come indicatore dello stato di peso forma. L'Indice di Massa Corporea è definito come:

$$\text{IMC} = \frac{\text{massa}}{\text{altezza}^2}$$

dove la massa è espressa in chilogrammi e l'altezza in metri. L'Organizzazione Mondiale della Sanità utilizza poi tabelle come la seguente per definire termini da "grave magrezza" fino a "super obesità".

*La survey su comportamenti e
conoscenze in campo alimentare*

D22

Può dirmi la sua statura? (in centimetri)

- Cm: _____
- Non sa / Non risponde

D23

Può dirmi il suo peso? (in kilogrammi)

- Kg: _____
- Non sa / Non risponde

vD22	vD22_txt	vD23	vD23_txt
1	171		184
1	162		158
1	1.70		174
1	160		185
1	180		170
1	172		168
1	1.60		175

- 1.70 non è un valore accettabile per l'altezza in centimetri
- verificiamo che il sistema CATI utilizzato consenta questo tipo di errore. Nel caso specifico, il campo è stato impostato come "carattere", perciò il sistema non ha effettuato una verifica di plausibilità del dato (che non viene considerato un numero)
- In questo caso l'errore è facilmente correggibile, in altri può dipendere da condizioni più complesse (ad esempio, valori che devono assommare ad un totale dato)

> x\$altezza

```
[1] 171 162 170 160 180 172 160 160 163 160 162 165 170 160 150 168 171 182 181 160 160 160 171 165 157 155 165 145 165 160 164 178 166 167 170 173 169
[38] 178 170 160 165 150 159 168 172 170 165 168 160 154 183 175 160 165 156 174 165 164 161 160 150 170 158 170 165 183 166 167 158 170 150 170 185 172
[75] 171 165 163 180 170 166 155 165 160 160 165 178 165 160 165 162 160 182 170 160 175 164 168 160 174 163 165 NA 162 160 160 155 152 160 160 175 165
[112] 152 170 164 160 170 175 158 162 163 172 160 164 152 168 163 160 171 160 155 175 170 162 175 157 155 160 166 175 170 165 170 150 161 155 162 167 NA
[149] 175 163 160 165 170 160 176 155 172 170 175 164 160 165 149 155 165 155 160 155 167 155 155 170 160 160 160 150 160 170 170 170 170 168 165 167
[186] 156 166 165 169 170 175 157 155 156 187 169 149 161 180 155 175 158 160 170 177 160 170 160 158 162 165 170 162 178 165 160 158 160 158 175 170 165
[223] 171 170 160 160 165 165 155 170 170 155 166 170 160 163 155 170 180 165 165 163 165 162 160 175 175 170 170 160 161 155 160 167 170 170 160 175 168
[260] 172 153 163 163 162 187 165 163 175 150 162 NA 185 160 160 165 160 152 172 158 163 178 158 170 156 160 158 172 170 170 160 155 162 170 160 157 165
[297] 165 170 160 150 158 158 168 170 155 180 160 164 164 NA 165 164 188 165 170 170 175 160 170 170 180 164 177 165 189 160 160 175 167 150 150 167 177
[334] 160 170 167 150 163 170 170 168 160 189 168 175 165 155 158 168 160 165 160 160 170 173 170 190 160 164 155 161 160 155 165 177 170 170 160 153 155
[371] 164 173 168 170 165 152 160 169 169 160 160 189 178 158 160 160 165 165 167 160 NA 154 165 173 152 160 162 172 168 162 163 155 165 172 165 180 167
[408] 170 162 178 175 167 158 184 168 165 170 172 173 170 155 155 174 165 148 165 171 168 163 165 165 NA 163 185 158 172 160 178 165 164 181 170 172 173
[445] 170 173 180 155 155 167 158 152 157 156 160 180 164 168 162 170 165 162 170 170 170 162 163 165 167 169 160 170 152 160 160 160 155 NA 178 180 154
[482] 160 170 160 155 165 161 151 180 154 162 165 155 184 160 160 185 164 160 165 162 160 150 171 170 175 170 160 165 152 150 160 162 170 168 170 160 183
[519] 185 153 160 184 153 170 180 155 183 170 175 165 170 160 160 187 165 160 150 150 170 184 162 179 170 155 171 165 173 150 172 165 168 169 170 160 165
[556] 165 162 170 162 167 167 162 170 162 NA 170 176 165 165 165 163 171 172 170 171 155 160 155 170 173 180 188 190 170 160 186 155 156 168 165 160 162
[593] 170 160 158 155 158 170 160 177 165 158 168 190 163 163 165 170 156 172 160 150 175 173 180 175 152 175 173 152 155 162 160 165 157 165 156 164 165
[630] 160 158 158 160 167 154 174 165 165 170 170 170 160 162 178 160 175 173 167 160 160 159 172 170 187 180 172 165 160 163 158 165 168 180 183 170 165
[667] 180 165 177 170 170 167 170 177 160 167 NA 170 184 170 170 185 162 150 160 168 165 160 170 165 170 168 162 168 NA 165 NA 175 NA 172 165 156 170
[704] 160 160 163 170 170 176 180 162 157 175 160 170 170 NA 163 180 156 160 173 155 158 180 164 165 168 165 178 160 155 168 150 160 176 170 168 168 170
[741] 158 184 170 180 180 162 160 177 170 173 170 165 163 172 174 173 162 188 185 170 168 155 178 182 NA 180 153 170 NA 180 160 175 160 165 168 176 170
[778] 165 174 175 163 175 175 181 165 180 175 167 170 190 175 180 180 174 180 167 180 182 170 175 155 176 180 175 180 172 165 181 176 170 170 183 173 175
[815] 176 172 162 173 179 170 165 175 164 180 160 165 175 165 170 168 170 170 170 170 165 164 180 172 165 169 170 160 160 179 170 180 177 160 167 163 175
[852] 168 183 163 170 158 168 NA 180 180 175 160 168 168 160 160 181 175 173 175 168 164 184 170 185 170 180 170 155 159 172 172 165 NA 165 182 172 165
[889] 158 163 165 160 186 172 152 166 172 160 160 160 NA 158 172 165 175 150 163 190 160 162 168 175 162 184 170 175 170 156 161 160 190 150 165 150 160
[926] 168 168 155 168 185 175 160 175 172 174 155 160 176 185 162 163 178 170 172 150 175 170 170 175 166 170 180 NA 175 177 163 170 170 173 170 170 170
[963] 165 162 165 170 163 170 170 170 161 178 170 165 180 167 178 172 185 175 182 180 166 170 160 180 180 170 189 170 164 180 175 184 182 153 175 175 175
[1000] 175 170 179 170 175
```

```
> summary(x$altezza)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
145.0  160.0   166.0   166.9  172.0   190.0   19.0

> summary(x$peso)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 35.00  58.00   66.00   66.97  75.00  158.00  30.00

> mean(x$altezza)
[1] NA
> mean(x$altezza,na.rm=T)
[1] 166.8904

> x$imc<-(x$peso/(x$altezza/100)^2)
> summary(x$imc)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 13.67  21.31   23.44   23.96  25.88   59.47   34.00
```

FREQUENCIES VARIABLES=vD22_txt
/ORDER=ANALYSIS.

Statistiche

Testo

N	Validi	1004
	Mancanti	0

Testo

	Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	19	1,9	1,9	1,9
1.50	4	,4	,4	2,3
1.52	1	,1	,1	2,4
1.55	6	,6	,6	3,0
1.56	2	,2	,2	3,2

FREQUENCIES VARIABLES=vD22_txt
/ORDER=ANALYSIS.

Statistiche

Testo

N	Validi	905
	Mancanti	99

Testo

	Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	145	,1	,1	,1
	148	,1	,1	,2
	149	,2	,2	,4
	150	1,9	2,1	2,5

1. Alcuni software statistici (es, SPSS) trattano i valori mancanti (missing, NA) semplicemente scartandoli nelle elaborazioni .
2. Questa pratica è accettabile solo se il meccanismo di generazione dei valori mancanti è MCAR (Missing completely at Random), cioè non dipende dalla variabile di interesse e da nessuna altra variabile
3. Se i missing sono solo una piccola percentuale (es, meno del 5%) comunque le distorsioni introdotte sono di scarsa importanza (cfr. J. Scheffer, *Dealing with Missing Data*, R.L.I.M.S, vol 3, 2002)
4. Nel nostro caso nella costruzione dell'IMC ho solo 34 NA su 1004 (3,4%) e, rispetto allo stratificazione adottata, non emergono elementi sistematici nel tasso di non risposta per il peso e per l'altezza

```
> summary(glm(indpeso~factor(vD0),data=df,family=binomial)
+ )

Call:
glm(formula = indpeso ~ factor(vD0), family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3203 -0.2561 -0.2355 -0.2309  2.8278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9444    0.5923  -4.971 6.67e-07 ***
factor(vD0)2  -0.4568    0.7060  -0.647  0.518
factor(vD0)3  -0.1801    0.7822  -0.230  0.818
factor(vD0)4  -0.6665    1.1738  -0.568  0.570
factor(vD0)5  -0.7932    0.9288  -0.854  0.393
factor(vD0)6  -0.6267    0.6820  -0.919  0.358
factor(vD0)7  -1.0352    0.9274  -1.116  0.264
factor(vD0)8  -0.5819    0.9304  -0.625  0.532
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 269.73  on 1003  degrees of freedom
Residual deviance: 267.74  on   996  degrees of freedom
```

- L'utilizzo dei dati completi comunque porta ad una diminuzione dei casi disponibili (34 nel nostro caso)
- conviene perciò imputare i dati mancanti
- si può imputare i dati mancanti in modo naif, ad esempio con un valore medio in ogni raggruppamento determinato dalla variabile di stratificazione (età, sesso)
- oppure si può utilizzare un più ampio set di informazioni (ad esempio, la relazione tra altezza e peso), con un approccio iterativo (poiché sono presenti NA sia nel peso che nell'altezza)
- sono disponibili software che producono queste imputazioni automaticamente, anche se è sempre opportuno analizzare prima i pattern dei valori mancanti
- S. Van Buuren, K Groothuis-Oudshoorn, MICE: multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 2010
- Da notare: **l'imputazione multipla** è l'approccio che raccoglie il consenso generalizzato degli statistici teorici, mentre la statistica ufficiale preferisce utilizzare **l'imputazione singola**, essenzialmente per la difficoltà di gestire ed elaborare archivi statistici multipli.

```
> summary(x$altezza)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.   NA's  
145.0  160.0  166.0  166.9  172.0  190.0  19.0
```

```
> summary(x$peso)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.   NA's  
35.00  58.00  66.00  66.97  75.00  158.00 30.00
```

```
> summary(df_imp$altezza)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.  
145.0  160.0  166.0  166.9  172.0  190.0
```

```
> summary(df_imp$peso)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.  
35.00  57.00  65.00  66.84  75.00  158.00
```

PISA 2003 – le competenze dei quindicenni, questionario studenti

Dodici variabili (ST29Q_i, con $i=1, 2, 3, 4, 5, 6$; ST339Q_i, con $i=1, 2, 3, 4, 5, 6$): tempo settimanale dedicato a sei attività di studio, per il complesso delle materie scolastiche e per matematica. Le attività considerate sono: compiti assegnati dagli insegnanti; corsi di recupero a scuola; corsi di potenziamento a scuola; ripetizioni con insegnante privato; frequenza di corsi al di fuori della scuola; altri tipi di studio.

Per ogni coppia di variabili si può calcolare la loro differenza, $ST29Q_i - ST33Q_i$, con $i=1, 2, 3, 4, 5, 6$, che misura le ore settimanali dedicate all'attività nelle materie diverse da matematica. Si noti che per ogni studente queste differenze devono logicamente essere maggiori o uguali a zero, perché le ore di matematica sono una parte delle ore nel complesso delle materie.

Relativamente al complesso dei paesi che hanno partecipato all'indagine, questo gruppo di variabili presenta due problemi:

1. un tasso di mancata risposta estremamente elevato (da un minimo del 15,2% ad un massimo del 31,9%)
2. la presenza di valori negativi, nel 2,7% dei casi validi, nella differenza tra ST29Q06 e ST33Q06.

Indici sintetici

- Indice sintetico delle “competenze nutrizionali” per verificare se l’IMC possa essere associato alle competenze nutrizionali dei soggetti
- punteggio ottenuto (somma delle risposte codificate 0,1) nelle risposte alle 12 domande (item) volte a verificare le conoscenze nutrizionali.
- verifica che i 12 item corrispondano ad un variabile latente unidimensionale (medesima scala)
- utilizzo dell’approccio non parametrico proposto da R. Mokken (scale di Mokken)
- Software che consente l’individuazione automatica di set di item appartenenti ad una medesima scala (L. A. van der Ark, Mokken Scale Analysis in R, Journal of Statistical Software, 2009)
- 9 item su 12 conservati nella scala, che ha buone proprietà statistiche

```
> table(df_imp$punteggioMokken)
```

```
 0    1    2    3    4    5    6    7    8    9  
33   18   24   49  120  209  184  168  114   85
```

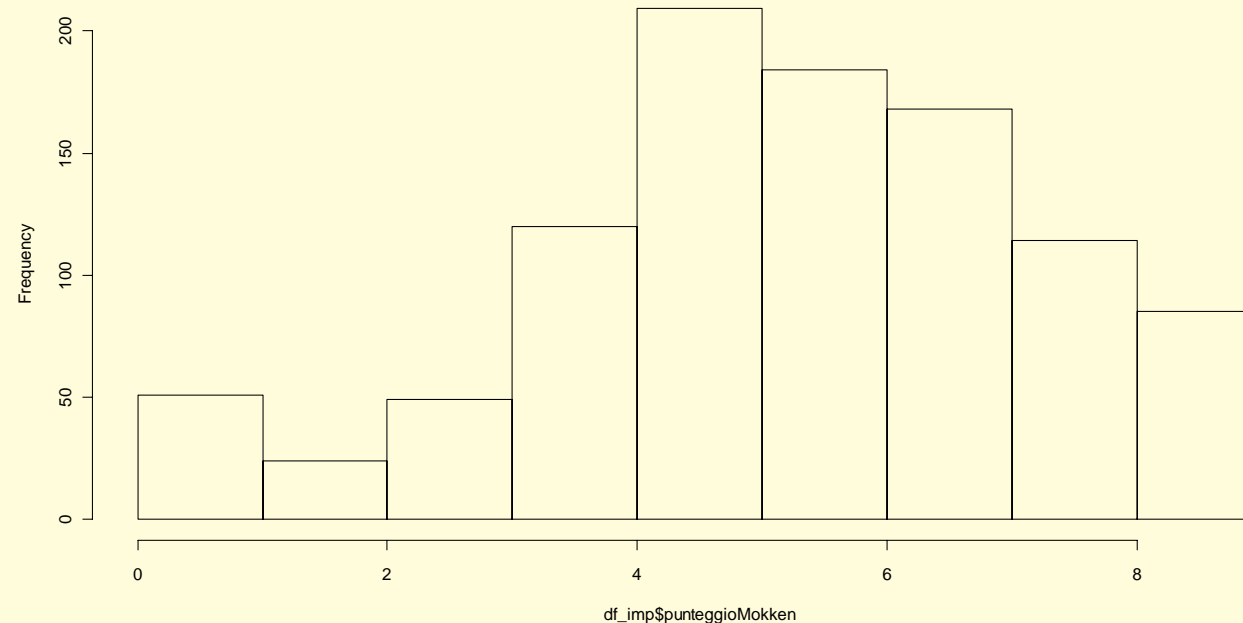
```
> cor(df_imp$punteggio,df_imp$imc)
```

```
[1] -0.06337624
```

```
> cor(df_imp$punteggioMokken,df_imp$punteggio)
```

```
[1] 0.9560066
```

Histogram of df_imp\$punteggioMokken



Ponderazioni, quando?

La survey su comportamenti e conoscenze in campo alimentare

**Tabella 6.1 – Universo: popolazione residente al 1 gennaio 2008
per classe di età e sesso (Regione Lombardia)**

Età	Maschi	Femmine	Totale
18-29	600.618	573.024	1.173.642
30-59	2.231.742	2.163.285	4.395.027
60-74	753.268	852.066	1.605.334
75 e oltre	305.822	570.253	876.075
Totale	3.891.450	4.158.628	8.050.078

Fonte: elaborazione IRER su dati ISTAT

Popolazione residente al 1 Gennaio 2009 per età e sesso e stato civile,
Lombardia

	Maschi	Femmine	Totale
18-29	597998	572270	1170268
30-59	2243577	2174921	4418498
60-74	768071	864359	1632430
75 e oltre	318267	582960	901227
Totale	3927913	4194510	8122423

*La survey su comportamenti e
conoscenze in campo alimentare*

Tabella 6.4 – Campione effettivo

Età	Maschi	Femmine	Totale
18-29	60	86	146
30-59	217	329	546
60-74	95	109	204
75 e oltre	38	70	108
Totale	410	594	1.004

n. n.

Tabella 6.3 – Campione teorico

Età	Maschi	Femmine	Totale
18-29	75	71	146
30-59	277	269	546
60-74	94	106	199
75 e oltre	38	70	108
Totale	483	517	1.000

Lo stimatore del totale di un carattere y () associato a questo piano di campionamento è dato da:

$$\hat{t}_y = \sum_{h=1}^L N_h \bar{y}_h$$

N_h $h = 1, 2, \dots, L$ sono le dimensioni degli strati nella popolazione obiettivo ().

($\sum_{h=1}^L N_h = N$). \bar{y}_h rappresenta la media campionaria calcolata sulle n_h unità statistiche campionate all'interno dello strato h ed è quindi definita

come $\bar{y}_h = n_h^{-1} \sum_{j=1}^{n_h} y_{hj}$.

Possiamo quindi scrivere

$$\hat{t}_y = \sum_{h=1}^L \sum_{j=1}^{n_h} \tilde{w}_h y_{hj}$$

$$\tilde{w}_h = \frac{N_h}{n_h}$$

- *Stima: trattamento ex-post della non risposta*

Si è ipotizzato un modello di non risposta estremamente semplice basato sull'ipotesi che la probabilità di risposta sia costante all'interno di ciascuno strato (in altre parole, le classi di aggiustamento per la non risposta coincidono con gli strati).

Ciò porta alla seguente correzione dei pesi base \tilde{W}_h .

Definito $n_{h,risp}$ il numero dei rispondenti effettivi tra le n_h unità campionate all'interno dello strato h , il peso aggiustato per non risposta è definito come

$$w_h = \frac{N_h}{n_{h,risp}}$$

e quindi lo stimatore effettivamente utilizzato per la stima di t_y sarà dato da:

$$t_y = \sum_{h=1}^L \sum_{j=1}^{n_h} w_h y_{hj} .$$

```
> print(svymean(~peso,df_design_08))
      mean      SE
peso 67.899 0.3726
> print(svymean(~peso,df_design_08_simple))
      mean      SE
peso 66.888 0.414
█
```

```
> print(svymean(~peso,df_design_09))
      mean      SE
peso 67.905 0.3726
```



BANCA D'ITALIA
EUROSISTEMA

Temi di Discussione

(Working Papers)

The use of survey weights in regression analysis

by Ivan Faiella

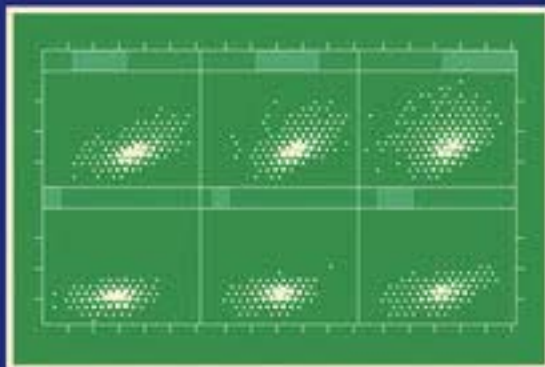
January 2010

number
739

WILEY SERIES IN SURVEY METHODOLOGY

Complex Surveys

A Guide to Analysis Using R



Thomas Lumley

Trovate dei materiali di approfondimento delle questioni trattate in questa presentazione sulla rete di istituto:

\\SERVERDATI2\PiazzaComune\PolicyAnalysis16giugno2010_gay